Fuzzy techniques for coffee flavour classification

Lotfi Khodja, Laurent Foulloy, Eric Benoit, Thierry Talou

Fuzzy techniques for coffee flavour classification

L. Khodja, L. Foulloy, E. Benoit, T. Talou *

Laboratoire d'Automatique et de Micro-Informatique Industrielle
LAMII/CESALP Université de Savoie
B.P. 806, 74016 Annecy cedex
FRANCE

* Laboratoire de Chimie Agro-Industrielle
Ecole Nationale Supérieure de Chimie de Toulouse
118, route de Narbonne - 31077 Toulouse cedex
FRANCE

E-mail : khodja, foulloy, benoit@univ-savoie.fr

Abstract

Coffee plants are grown in different regions. Each region produces its own flavours and characteristics. Robusta and arabica coffees differ from each other, but also each category develops a great number of tastes. The blending of two coffee beans offers then an infinite choice of aromatic properties. This paper deals with the differentiation of pure and mixed samples of commercialized coffees. Fuzzy c-means and k-nearest neighbours algorithms are used to discriminate pure arabica and robusta samples by treating the initial feature vectors. When samples sets including mixtures are analysed, these techniques fail to differentiate efficiently this category of samples. A data processing by means of a discriminant analysis provides two new features which are used for classification purposes using the k-nearest neighbours algorithm.

Introduction

There are many species of coffee beans, but only two are commercially traded : robusta and arabica which are produced by two coffee tree species : *Coffea Arabica* and *Coffea Camphora Robusta*. Their differences lie in their seeds and the soil and altitude at which they are grown. Arabica plants are less plentiful than robustas because they are often grown at higher altitudes. Arabica beans are flavorful and sell at higher prices because of their quality and limited availability. Since the cultivation of coffee trees and the processing techniques vary from country to country, the flavours also vary. Besides, the different possibilities of coffee beans blending create an infinite number of options and flavours. Hence the necessity to dispose of measurement and data processing techniques allowing to differentiate different coffee samples depending on their aromatic properties.

More generally, the qualification and quantification of aroma volatiles emitted by flavourings, plants extracts or flavoured industrial products are two important factors of quality control methodology used by aromatic and food industries. Physico-chemical techniques and sensory analysis are two classical methods used for this purpose. But these techniques are expensive and time consuming. Recently, considerable interest has arisen in the use of array of gas sensors together with an associated pattern recognition technique to measure, differentiate, and identify complex mixtures of volatile compounds. The sensors used are of MOS (Metal Oxide Sensor) type, and are composed of tin dioxide deposit on an alumina ceramic tube with a heat coil inside. The principle of the detection of such an apparatus, labelled "electronic nose ", is based on the reversible electrical resistance changes of the sensing elements in presence of volatiles. The adsorption kinetics is of several seconds, and the adsorption is proportionnal to volatiles concentration in the atmosphere [1].

Fuzzy logic based techniques have proven their efficiency in performing symbolic description of measurement [2], [3]. Fuzzy clustering techniques may be used to give an objective description of human visual sense, for example, by differentiating colours [4]. In this paper we are concerned with the discrimination of different and complex mixtures of coffee samples. We use the fuzzy-c-means and the k-nearest neighbours algorithms in order to classify these samples. The c-means algorithm computes a clustering by minimization of the within-group variance. The k-nearest neighbours algorithm permits to assign an object to classes for which some prototypes are predefined. These techniques allow us to efficiently differentiate pure samples of the Arabica and Robusta families, but the misclassification rate for mixtures is too high. These techniques are based on a distance in the space of the features which permits to measure how far two samples are from each other. In our case the data belong to a 5 dimensionnal space and are provided by 5 sensors which measure the concentration of volatiles in the atmosphere. In the space of data, the points corresponding to pure arabica and pure robusta

samples can be majoritarily put into two convex regions. But when samples sets including mixtures are analysed, these techniques fail to differentiate efficiently the mixed samples. Actually, the points corresponding to the mixtures do not lie into an intermediary region between arabicas and robustas, but some of them can be very close to pure arabica points and some other to pure robusta points. Not only can arabica (robusta) coffees differ very much from each other, but also, as stated above, blend possibilities create an infinite number of flavours. The analysed samples are taken from commercialized coffees whose origins and processing history are ignored. Mixtures with the same arabica and robusta concentrations may be closer to pure arabica or pure robusta flavours depending on the aromatic properties of the mixed coffees. This is confirmed by a principal component analysis which shows that the mixtures do not belong to an isolable area of the data space. A processing of data by means of a discriminant analysis, followed by the k-nearest algorithms permits to differentiate much more efficiently the samples of coffee to analyse.

1- Measurements

45 different samples of commercialized, vacuum-packed, ground coffee have been analysed. These samples are of three categories : arabica (20 samples), robusta (10), and mixtures (15). Five different volatiles sensors have been used. These sensors are placed into a 600 millilitres measurement cell. 5 grams of coffee pouder are placed into a 125 millilitres recipient, and a 5 volts voltage is applied to the heat coil. The coffee pouder generates then volatiles during a 5 minutes diffusing time, before taking a 50 millilitres gas sample by the means of a gas syringe. Before introducing the gas into the measurement cell, the initial resistances of the sensors are measured (R_{air}). The injection of the volatiles makes the resistances decrease (R_{gas}). The variations of the electric resistances are then analysed. A computer is used for the experiment control and the electric signals acquisition via 5 A/D converters. Before analysing a new sample, the measurement cell is cleaned by means of a compressed air circulation.

2- Two fuzzy classification techniques

2.1- Introduction

The concept of a fuzzy set deals with the representation of classes whose boundaries are not quite determined. Instead of the binary characteristic functions associated with the usual “hard” sets, a fuzzy set is fully described by its membership function which takes all the possible degrees intermediary between 0 and 1. In the field of cluster analysis, it may be more realistic for describing a data set to look for fuzzy clusters when some clusters are not well separated.

2.2- Classification by the fuzzy c-means algorithm (FCM)

The fuzzy-c-means (FCM) clustering algorithm [5] the fuzzy equivalent of the nearest mean hard clustering algorithm [6]. Data are supposed to be structured into n vectors whose dimension is p : X_j ; j = 1,2,...,n ; each vector characterizes an object described with p attributes. We assume here that the number of clusters is known. For this preselected number c, the FCM algorithm produces c vectors which represent the cluster centers and for each data point c membership values which measure the similarity of the data points to each of the cluster centers. Let u_{ij} be the membership value of the vector X_j to the cluster i describing how close X_j is to this cluster's center C_i. The classification is obtained by minimizing the following objective function with respect to the memberships $\{u_{ij}\}$ and the cluster centers $\{C_i\}$:

$$\sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m d^2(X_j, C_i)$$

where d is a distance in the space of data. The value of the fuzzy index m tunes the degree of fuzziness of the clustering. The membership values indicate how well the point has been classified. When the input is close to a particular center, the membership value to the corresponding class is close to one. In the case of uniformly low memberships the point can not be classified clearly. First the memberships are given guessed initial values such that , for $j = 1, 2, ...,n$:

$$\sum_{i=1}^{c} u_{ij} = 1 .$$

The following iterative procedure converges to a minimum of the objective function [5]:

- compute the cluster centers :

$$C_i = \frac{\sum_{j=1}^{n} (u_{ij})^m \cdot X_j}{\sum_{j=1}^{n} (u_{ij})^m}$$

- update the fuzzy membership functions :

$$u_{ij} = \frac{1/(d(X_j, C_i))^{2/(m-1)}}{\sum_{i=1}^{c} 1/(d(X_j, C_i))^{2/(m-1)}}$$

2.3- Classification by the fuzzy k-nearest neighbors algorithm (KNN)

A fuzzy k-nearest neighbors algorithm was proposed

by Keller & al [7]. The conventional KNN classification method assigns each input to one of the possible classes. First the k-nearest neighbors are found. Then the input is assigned to the class which includes the majority of the neighbors. In the fuzzy KNN classifier, the second step consists in assigning to the input a membership degree to each class. Let x be an input, $u_i(x)$ its membership to the i^{th} class and w_{ij} the one of the j^{th} neighbor z_j, then :

$$u_i = \frac{\sum_{j=1}^{K} w_{ij} \cdot \left(\frac{1}{\|x - z_j\|}\right)^{2/(m-1)}}{\sum_{j=1}^{K} \left(\frac{1}{\|x - z_j\|}\right)^{2/(m-1)}}$$

where m is an adjustable parameter which tunes the weighting effect of the distance. When m approaches 1 the nearest samples have much more effect on the membership of the input. The "defuzzification" consists in assigning the input to the class to which it has the highest membership degree.

3- Differentiation of coffee samples

3.1- Arabica/Robusta discrimination

The measurements are made on 45 different commercialized ground coffees, usually sold in supermarkets. Twenty of these samples are pure arabica coffees, ten are pure robusta ones, and fifteen are mixtures. Table 1 gives the measurements obtained with five different sensors (TGS 825, TGS 800, TGS 824, TGS 822 and TGS 812) for the first 30 samples whose the first 20 are pure arabica samples (A) and the the last 10 are pure robusta ones (R) as shown in the last column of this table.

Table 1 : Measurements

TGS825	TGS800	TGS824	TGS822	TGS812	
47.04	74.56	48.09	109.04	69.86	A
52.46	81.28	55.41	111.14	73.54	A
44.32	63.57	43.73	104.30	60.74	A
44.62	65.89	40.68	100.71	62.51	A
44.36	61.14	39.59	98.76	58.37	A
48.20	62.00	22.15	98.64	48.50	A
43.27	62.99	41.56	98.64	59.54	A
47.46	74.30	48.80	100.74	66.03	A
54.69	80.28	52.55	114.34	73.63	A
52.96	78.87	53.93	106.61	73.25	A
46.74	76.10	47.71	99.56	65.38	A
49.73	71.71	50.65	105.63	69.44	A
47.04	72.41	50.15	99.16	63.61	A
44.25	61.91	43.25	102.00	62.13	A
42.60	65.89	41.51	100.31	61.00	A
45.67	70.94	50.85	105.08	68.41	A
44.17	63.43	41.88	99.77	60.10	A
50.21	85.53	56.64	108.83	74.27	A
44.53	68.57	45.26	98.64	64.93	A
51.28	70.79	32.42	104.09	57.15	A
30.33	40.34	16.43	63.41	38.84	R
48.70	70.79	27.62	100.71	55.88	R
41.80	57.51	26.24	92.04	50.42	R
35.53	53.12	29.53	85.91	53.39	R
38.59	61.14	37.61	89.93	59.18	R
43.75	61.19	35.13	100.97	59.50	R
42.15	58.09	21.33	88.25	47.50	R
46.74	66.66	37.81	100.71	64.90	R
41.78	61.39	34.04	91.66	56.72	R
44.24	64.02	21.06	94.86	49.71	R

The results obtained with the set of pure coffees when using FCM are given in the table below where m was chosen equal to 2. The last two columns contain the expexted (e) and the computed class (c) of each sample (Arabica or Robusta).

Table 2 : membership degrees (FCM)

sample	u(R)	u(A)	e	c
1	0.038	0.961	A	A
2	0.120	0.879	A	A
3	0.232	0.767	A	A
4	0.260	0.739	A	A
5 (!)	0.604	0.395	A	R
6 (!)	0.880	0.119	A	R
7 (*)	0.459	0.540	A	?
8	0.017	0.982	A	A
9	0.132	0.867	A	A
10	0.091	0.908	A	A
11	0.039	0.960	A	A
12	0.024	0.975	A	A
13	0.044	0.955	A	A
14	0.288	0.711	A	A
15	0.307	0.692	A	A
16	0.021	0.978	A	A
17	0.380	0.619	A	A
18	0.140	0.859	A	A
19	0.094	0.905	A	A
20 (*)	0.474	0.525	A	?
21	0.729	0.270	R	R
22	0.684	0.315	R	R
23	0.980	0.019	R	R
24	0.907	0.092	R	R

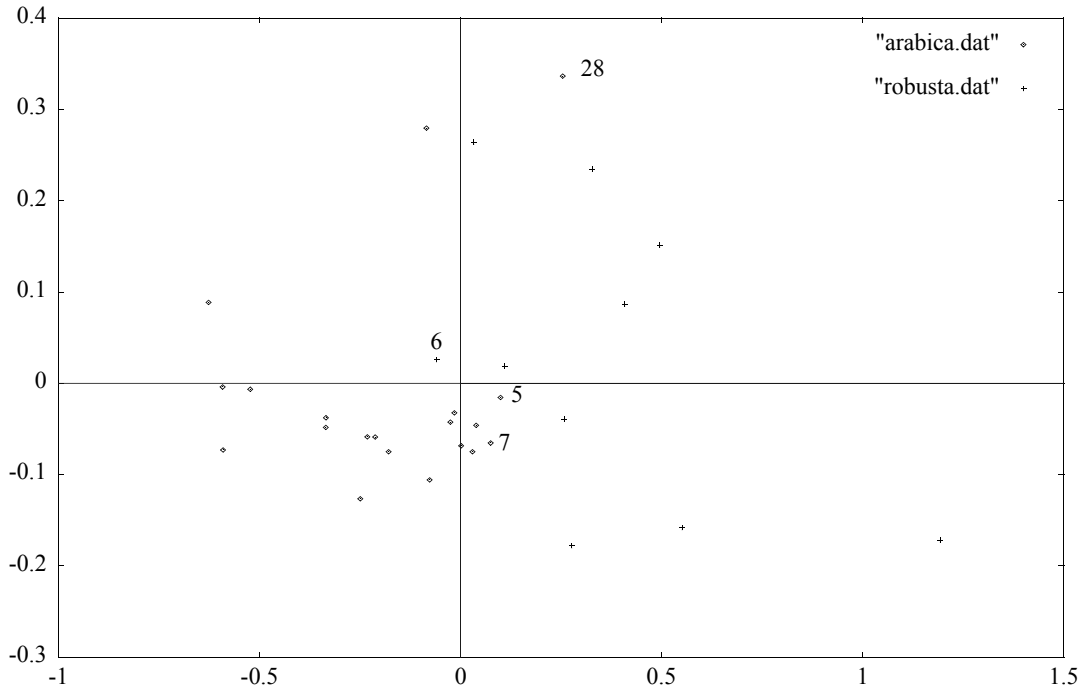

Figure 1 : projections of the data points on the plane of the two first factorial axes (PCA)

sample	u(R)	u(A)	e	c
25	0.805	0.194	R	R
26	0.687	0.312	R	R
27	0.930	0.069	R	R
28 (!)	0.277	0.722	R	A
29	0.933	0.066	R	R
30	0.917	0.082	R	R

Three samples have not been assigned to the class to which they are supposed to belong (marked with the (!) in the table). The misclassification rate is here equal to 10%. Two samples whose the highest membership degrees are smaller than 0.6 are marked with the character (*). The results obtained with KNN where 3 prototypes for each class have been defined lead us to the same conclusions. These results are presented in Table 3, where m_r is the misclassification rate.

Table 3 : results for pure samples

misclassified	uncertain	total	m_r
3	2	30	10%

The projections of the data points on the plane of the first two factorial axes computed by means of a principal component analysis (which represent over 95% of the inertia) show that the misclassified samples are obviously not located in their expected category area (Figure 1).

3.2- Data sets including mixtures

When data sets including mixtures are analysed, we do not differentiate efficiently the pure samples from the mixed ones. Either with FCM or with KNN, we observe that the clusters obtained with the pure 30 samples are no longer well differentiated. Samples of the arabica category for example are identified as robusta, and some other are identified as mixtures, and vice versa. In Table 4, XY represents the number of samples supposed to be of category X and which are assigned with KNN to category Y, (X, Y = A (arabica), R (robusta), M (mixtures)). The third column gives the numbers that would be obtained in an ideal case. Such results show that the points representing the samples in the data space are distributed such that there is no convex region containing a majority of the mixed samples.

Table 4 : results (KNN)

AA	18	20
AR	2	
AM	0	
RA	1	
RR	8	10
RM	1	
MA	7	
MR	4	
MM	4	15

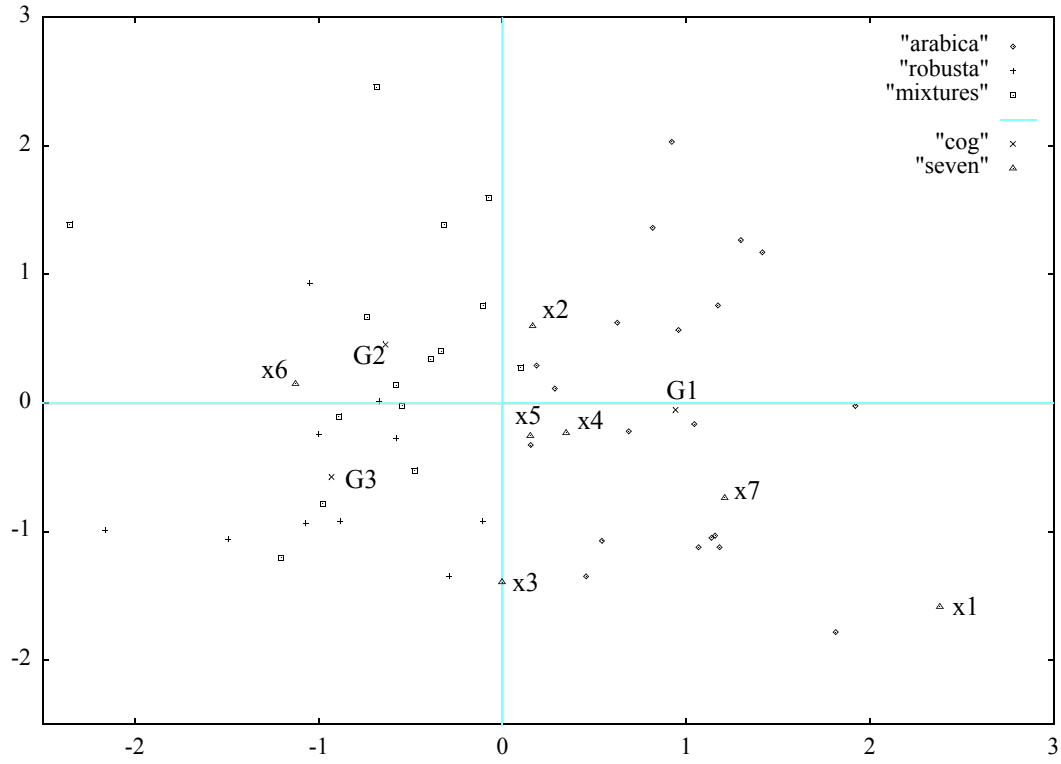

Figure 2 : Projections of the data points obtained by a discriminant analysis.

3.3- Statistical processing of data

Seven new samples are added and the new set of 52 samples is analysed by means of a discriminant analysis. Table 5 gives the arabica concentration of these seven samples. Samples x_6 and x_7 are two mixtures with unknown concentrations.

Table 5 : additionnal mixtures

sample	arabica concentration
x_1	100%
x_2	70%
x_3	0%
x_4	100%
x_5	100%
x_6	? %
x_7	? %

Figure 2 shows the projections of the 52 samples on the plane of the first two factorial axes computed by means of a discriminant analysis. The centers of gravity of the 3 categories (G1 for arabica, G2 for robusta and G3 for mixtures) are also shown. Notice that all the arabica points are located in the right hand half of the figure. The sample x_7 is projected in the arabica area while it is expected to belong to mixtures family. Table 6 gives the number of samples assigned to each class when applying KNN to the projected points corresponding to the initial 45 samples (AR is the number of arabica samples assigned to the robusta class). The number of misclassified samples when using discriminant analysis (with DA) is lower than the one obtained when analysing the raw data (without DA).

Table 6 : comparative results

	without DA	with DA
AA	18	19
AR	2	0
AM	0	1
RA	1	0
RR	8	7
RM	1	3
MA	7	0
MR	4	4
MM	4	11

Table 7 gives the results obtained for the seven additionnal samples when using KNN with all the 45 initial samples defined as prototypes and a number of k-nearest neighbors equal to 10.

Table 7 : classification of the additionnal mixtures

S	arabica	robusta	mixtures
1	1.000	0.000	0.000
2	0.424	0.000	0.575
3	0.311	0.626	0.062
4	0.940	0.017	0.041
5	0.887	0.028	0.084
6	0.000	0.413	0.586
7	1.000	0.000	0.000

Only x_7 is not assigned to the class to wich it is supposed to belong.

4- Discussion

The application described in this paper consists in defferentiating commercialized coffee samples. Fuzzy clustering is used to perform a symbolic description of the numerical data provided by flavours sensors. These data are in the form of 5-dimensionnal feature vectors. FCM and KNN allow to differentiate pure samples (arabica and robusta) in the original space of data. These techniques, however, do not perform good clusterings when analysing data which include mixed samples. As it may be expected, it is much easier to differentiate gourmet coffees (fine and expensive) than to differentiate poor quality ones. The performances should be taken in the context of this application. The difficulty lies in the fact that certain mixtures are very close to pure arabica samples, while some others are very close to pure robusta samples. Both FCM and KNN techniques compute clusterings based on the similarity (or dissimilarity) between two samples which is measured by a distance. Euclidian distance in the feature space leads to highly overlapping clusters. When applying a discriminant analysis and projecting the data points on the first factorial plane, we have 2 new more discriminating features and the results obtained KNN are much better. KNN has the advantage of being an automatic clustering technique, though when discriminating mixed samples here, we first statistically process the available numerical data.

Conclusion

In this paper we presented some results concerning the differentiation of coffee aromatic properties. We easily discriminated pure arabica and pure robusta coffee powders using two well known fuzzy clustering techniques : the fuzzy-c means and the the k-nearest neighbours algorithms. We dwelled on the difficulty to differentiate sets of samples including mixtures when using these same techniques in the space of the original features. This is inherent to the characteristics of coffee varieties whose mixing possibilities develop an infinite number of flavours. When the initial data points are projected on the plane of the first two factorial axes computed by means of a discriminant analysis, the fuzzy k-nearest neighbours technique permits to analyse much better data sets containing mixed samples.

References

[1] B. Bourrounet, T. Talou and A. Gaset, *Sensors and Actuarors* B, Vol. 27, N° 1-3, 1995, pp. 250-254.

[2] L. Foulloy, "Fuzzy Sensors for Fuzzy Control", *Proc. Int. Conf. IPMU*, Palma, Spain, July 1992, pp. 759-768, also in *Int. J. of Uncertainty, Fuzziness and Knowledge Based Systems*, Vol.2, No.1, 1994, pp. 55-66.

[3] G. Mauris, E. Benoit, L. Foulloy, "Fuzzy Symbolic Sensors : from Concept to Applications", *Measurement* 12 (1994), 357-384.

[4] L. Khodja, L. Foulloy, E. Benoit, “Some Fuzzy Clustering Techniques and Applications”, *Third International Applied Statistics in Industry Conference*, Dallas, 5-7 June 1995.

[5] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms". Plenum, New York, 1981.

[6] R. O Duda and P. E. Hart, "Pattern Classification and Scene Analysis". New York : Wiley, 1973.

[7] James M. Keller, Michael R. Gray, and James A. Givens, JR. "A Fuzzy K-Nearest Neighbor Algorithm", *IEEE Trans. Syst., Man, Cybern*., vol. SMC-15, No. 4, pp 580-585, July/August 1985.